LA-UR-05-9158

Title:	SEARCH ENGINE COVERAGE OF THE OAI-PMH CORPUS
Author(s):	Frank McCown, Xiaoming Liu, Michael L. Nelson, Mohammad Zubair
Submitted to:	IEEE Internet Computing

Search Engine Coverage of the OAI-PMH Corpus

Frank McCown, Xiaoming Liu, Michael L. Nelson, Mohammad Zubair

Abstract

The major search engines are competing to index as much of the Web as possible. Having indexed much of the surface Web, search engines are now using a variety of approaches to index the deep Web. At the same time, institutional repositories and digital libraries are adopting the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to expose their holdings, some of which are indexed by search engines and some of which are not. To determine how much of the current OAI-PMH corpus search engines index, we harvested nearly 10M records from 776 OAI-PMH repositories. From these records we extracted 3.3M unique resource identifiers and then conducted searches on samples from this collection. Of this OAI-PMH corpus, Yahoo indexed 65%, followed by Google (44%) and MSN (7%). Twenty-one percent of the resources were not indexed by any of the three search engines.

Introduction

Google, Yahoo, MSN, and other search engines are crawling and indexing as much content as possible to establish market preeminence. Meanwhile academic and research institutions are expending enormous effort to digitize their collection of theses, white papers, technical reports, maps, images, and historical documents to make them available in institutional repositories or digital libraries (DLs). These DLs represent significant institutional investment, yet their resources often remain hidden in the deep Web [1], the part of the Web that is typically hidden from Web search engines. DL maintainers wanting to expose their DLs to search engines have had to develop crawler-friendly web pages in an effort to coax web crawlers to index their websites. Unfortunately web crawlers will sometimes stop short of indexing all the pages for various reasons or may skip over some available content. Wishing to avoid the extra server load and network bandwidth costs associated with search engine crawlers, some DLs use the robots exclusion protocol (robots.txt) to protect their holdings from being crawled.

Not all deep Web resources are inaccessible. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is used by many DLs to expose metadata about their contents (see OAI-PMH sidebar). By issuing an OAI-PMH request, an XML-encoded list of all the metadata records held by a DL will be returned. Frequently this metadata contains a URI for the resource on the Web or to a web page from which the resource may be obtained. These web resources may remain hidden from web crawlers if there are no links in the indexable Web that point to these resources or if they are protected by robots.txt.

Figure 1 shows a web page (on the right) that has been indexed by Google through web crawling. This same page could also have been discovered using OAI-PMH by extracting the Dublin Core (DC) identifier from the OAI record (bottom left).

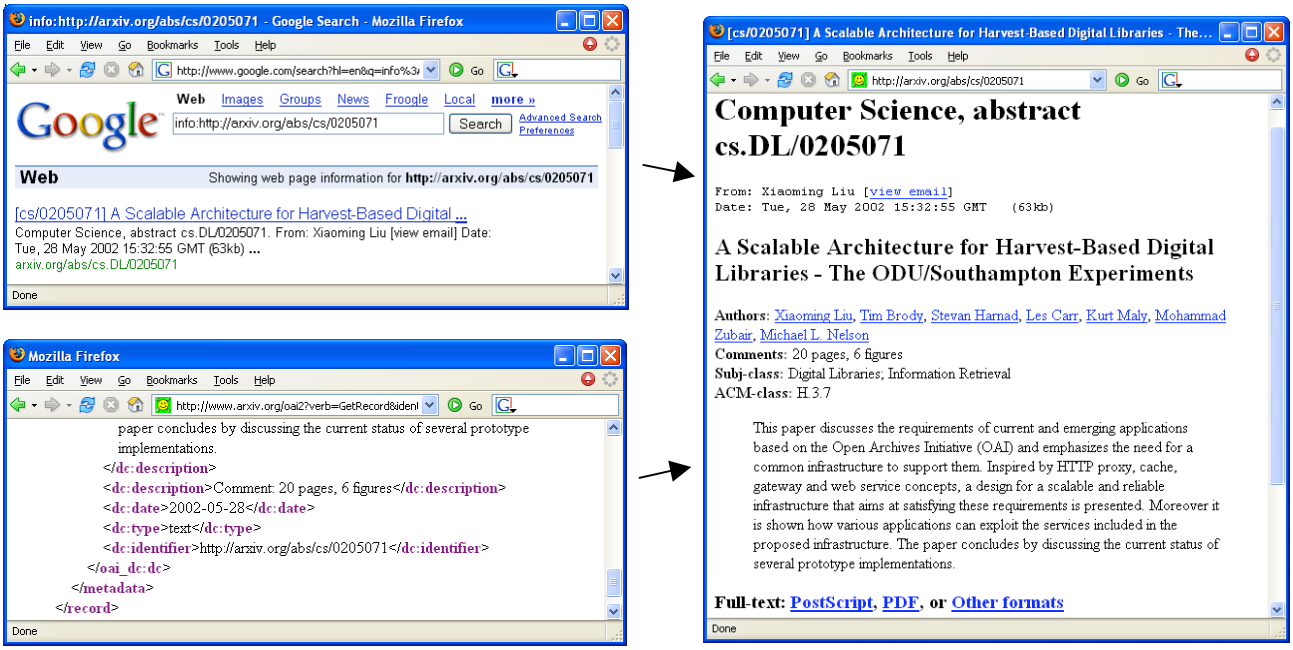

Figure 1 – Web page indexed by Google and pointed to in an OAI-PMH response

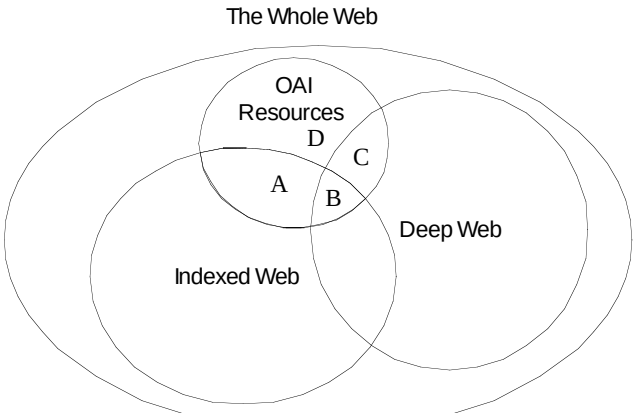

Figure 2 – Relationship between the indexed Web, deep Web and OAI resources

Figure 2 illustrates the partitioning of the Web and those resources pointed to by OAI-PMH (using DC identifiers) that are Web-accessible. The OAI Resources are divided into 4 parts:

A. Resources that have been indexed by search engines using crawling
B. Resources that have been indexed by search engines using OAIster, sitemaps and other techniques (see "Harvesting the Deep Web" sidebar)
C. Resources that are not accessible on the surface Web
D. Resources that are accessible on the surface Web and have not yet been found by crawling.

To measure search engine coverage of the OAI-PMH corpus (A and B), we collected a list of OAI-PMH repositories from four popular registries and harvested all their records. We then extracted URLs from the DC identifier fields of the harvested records and queried Google, MSN, and Yahoo to see if the URLs had been indexed.

Methodology

There is no central registry for OAI repositories. Typically repository administrators will register their repositories with one or more of the four well known registries [URLs 1-4]. From the four registries we collected 776 unique OAI repositories. We know of additional unregistered repositories, but we focus only on those that anyone could discover.

In June 2005 we harvested 9,843,451 DC records from 475 of the 776 OAI repositories (61%). The repositories that returned incomplete or erroneous responses may have been test repositories that have not yet been populated (pre-test) or are deprecated (post-test). 417 of the repositories (88%) returned at least one record, and 406 of the repositories (86%) returned at least one record with a DC identifier. Because we are interested in DC identifiers, we focus on this group of 406 repositories and call it the *repository corpus* (RC).

From the RC we extracted 5,575,375 DC identifiers; 4,376,271 (79%) were unique. The overlap of identifiers is common since some repositories harvest records from other repositories, and a single resource can be described by multiple metadata records. From the 5.6M DC identifiers we extracted 4,042,026 identifiers (73%) that began with "http://", "https://", and "www.". We refer to these as *resource identifiers*. 3,269,002 of the 4M resource identifiers (81%) were unique. The number of resource identifiers per repository varied greatly (Table 1).

Group	Resource IDs per Repository	Total Repositories		Total Resource IDs	
	0	87	18.3%	0	0.0%
1	1 – 999	267	56.2%	48,376	1.2%
2	1,000 – 9,999	80	16.8%	257,559	6.4%
3	10,000 – 49,999	28	5.9%	641,447	15.9%
4	50,000 – 99,999	6	1.3%	457,863	11.3%
5	100,000+	7	1.5%	2,636,781	65.2%
Totals		475	100.0%	4,042,026	100.0%

Table 1 – Grouping of repositories by number of resource identifiers

We sampled from the 4M+ resource identifiers based on three cross-sections:

1. the set of all unique DC identifiers
2. repositories based on the number of DC identifiers they contained
3. a representative group of 10 repositories

In each case we randomly chose 1000 resource identifiers to maintain at least a 95% confidence level (± 1%). Details of what resource identifiers were sampled are presented in the Results section. For each of these samples we ran tests to determine if the sampled resource identifiers pointed to actual content on the Web, and how many of these resource identifiers were indexed by Google, MSN, and Yahoo. We chose to use Google, MSN, and Yahoo as our test search engines since they are widely known and are three of the largest search engines [2]. All three search engines provide a mechanism to determine if an arbitrary URL is indexed or not.

Handles [5], Digital Object Identifiers (DOIs) [3] and Persistent URLs (PURLs) [4] made up 4.9% of the unique resource identifiers we extracted. When accessing these types of resource identifiers, an http 302 (found) is returned along with a temporary URL where the resource can be found. Because the returned URLs are temporary, we used the resource identifiers, not the temporary URLs, when querying the search engines.

Results

RC Results

We queried Google, MSN, and Yahoo for 1000 randomly selected, unique resource identifiers from the entire RC. Yahoo had the highest coverage with 65%. Google had indexed 44%, and MSN only 7%. 21% of the resource identifiers were not indexed by any of the search engines. Almost all 73 resource identifiers indexed by MSN were indexed by Google (96%), and many were indexed by Yahoo (78%). 69% (302/437) of the resource identifiers indexed by Google were also indexed by Yahoo.

We performed an http GET on each resource identifier and found that most were accessible: 96% of the requests resulted in an http 200 (OK) response. Only 3% resulted in a 404 (not found) response, and 1% resulted in some other response. Upon examination of the MIME types, we found 94% of the resources were "text/html", 3% were "application/pdf" and 2% were "text/plain". Only 1% of the resource identifiers were images ("image/gif" and "image/jpeg"). Similar responses and MIME types were obtained for the resource identifiers in other samples.

Repository Size Results

We randomly selected 1000 unique resource identifiers from each of the five repository groupings in Table 1. The results are shown in Figure 3. MSN faired poorly in all groupings, the highest being 35% of size group 1. All 3 indexed fewer resources as the size groups increased from 1 to 4 except the noticeable improvement made by Google at size group 4. Yahoo performed the best with the largest repositories (group 5) which is likely due to the OAIster agreement [6]; 6 of the 7 repositories in this group have been harvested by OAIster.

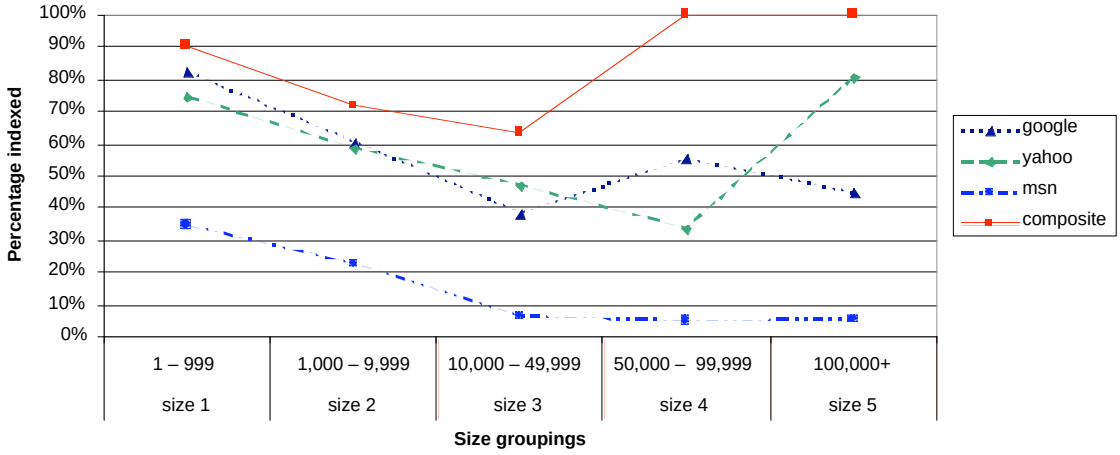

Figure 3 – Percentage of indexed resource identifiers for all five groupings

Representative Group Results

The first 2 sample statistics deal with the RC as a whole, but we also wanted to measure the success rate of indexing resource identifiers from a selective group from the RC. We manually selected 10 repositories for a cross section of content, geographical location, size of holdings, and implementation technique. From each of these repositories we randomly selected 1000 unique resource identifiers (or all if 1000 were not available).

Repository	Subject	Total records	Total DC identifiers	Sampled resource IDs	G %	M %	Y %	Comp %
Archeologia e Calcolatori - Published Articles	Institutional repository	383	20	20	95.0	90.0	100.0	100.0
UCT CS Research Document Archive	Institutional repository	100	100	100	100.0	40.0	100.0	100.0
Universidad de Chile - Tesis Electronicas	Theses and dissertations	407	407	407	85.0	29.7	94.6	98.8
Perseus Digital Library	Humanities	1,652	1,652	1000	9.2	0.0	99.6	99.6
Georgia Tech's Institutional Repository	Institutional repository	4,622	4,622	1000	71.5	0.2	0.0	71.7
BioMed Central	Publisher	17,357	17,357	1000	82.0	30.7	69.4	96.8
CERN Document Server	Institutional repository	38,939	62,654	1000	74.5	3.9	27.4	75.0
National Institute of Informatics Metadata DB	Institutional repository	76,582	89,238	1000	97.5	17.3	24.4	99.0
Library of Congress OAI Repository 1	US Government	191,664	191,663	1000	1.3	0.4	99.1	99.1
NDLTD Union Catalog	Theses and dissertations	199,099	212,799	1000	67.8	8.0	61.4	81.1
Average		53,081	58,051	753	68.4	22.0	67.6	92.1

Table 2 – 10 representative repositories from the RC

Table 2 lists the 10 repositories selected and the percentage of resource identifiers indexed by each search engine and the composite. Between Google and Yahoo there is no clear winner: Google outperformed Yahoo in 5 of the repositories, and Yahoo outperformed Google in 4 of the repositories. MSN was outperformed in all repositories and only performed well in indexing the smallest repository.

Observations

Recent published estimates show that Google has indexed more of the Web than Yahoo, and Yahoo has indexed more than MSN [2]. Our experiment reveals that Yahoo has indexed more resource identifiers from the entire RC than Google. Yahoo and Google performed similarly when indexing resource identifiers from repositories of different sizes except the largest size from which Yahoo performed the best. The largest repositories (size 4 and 5) appear to have the highest composite coverage despite the low coverage offered by MSN.

Perhaps the most interesting difference between Yahoo and Google can be seen in what they did and did not index in the 10 representative repositories. Yahoo performed significantly better than Google in indexing resource identifiers from Perseus Digital Library and Library of Congress (LC). In both cases Yahoo had indexed close to 100% of the resource identifiers compared to less than 2% for Google. Yahoo's success in indexing resources from these repositories is likely due to the arrangement between OAIster and Yahoo which included these two repositories.

Manual examination of the Perseus and LC websites revealed that a crawler could have found many of the resource identifiers on the Perseus site. The LC resource identifiers use handles which would not have been found by crawling, but the URLs pointed to by the handles could have been found with crawling. On both websites we found a robots.txt file that protected these resources from being crawled. We examined the Internet Archive for older versions of the robotos.txt and found that the same URL patterns had been protected for several years. Because popular search engines like Google, MSN, and Yahoo generally respect the robots exclusion and will not crawl content protected by robots.txt, it is likely that Yahoo indexed the Perseus and LC resource identifiers (and possibly metadata) but did not crawl and index the actual web pages pointed to by the resource identifiers. This is further supported by the fact that we performed some manual searches and were unable to find specific content from the resource identifiers' web pages in Yahoo.

Google outperformed Yahoo indexing resource identifiers from the CERN Document Server (75% vs. 27%) and Georgia Tech's institutional repository (72% vs. 0%). CERN is harvested by OAIster, but Georgia Tech's repository is not. It is possible that many of the CERN resource identifiers were not harvested by OAIster prior to the agreement with Yahoo.

A majority (64%) of the CERN resource identifiers pointed to content that could have been found through web crawling but was protected by a robots.txt file. Archived versions of the robots.txt go back 2 years and protect the same URLs from being crawled. Since neither Yahoo nor MSN had indexed any of these resource identifiers, perhaps they were crawled by special permission or using the Google Sitemap. A cursory look at the URLs two months after our tests showed that many of the URLs had fallen out of the Google index.

All the Georgia Tech resource identifiers use handles that we were unable to resolve. We performed http GET requests on these handles 3 times during June and July 2005 and were returned a "Cannot connect to server" error page for every URL. Apparently the handles pointed to a web server that was unable to return valid content. Google appears to be keeping these URLs indexed despite the error pages.

The OAIster agreement and the robots exclusion seem to have played a large role in how well Google and Yahoo performed in indexing resource identifiers from several repositories. It is likely that these repositories used robots.txt to reduce the web server load caused by robots crawling the dynamically generated pages. Since OAI-PMH offers incremental access by datestamp, new and modified resources could be discovered with less server overhead.

Returning to Figure 1, we measured:

A + B = 2.6M resource identifiers
C + D = 700K resource identifiers

We had hoped to solve for A and B directly, but we found that the search engines would not reliably report "backlinks", and therefore we could not be sure how a search engine discovered a resource. Solving for C and D directly is difficult because we cannot determine if an arbitrary URL is discoverable (but not yet discovered) on the surface Web or if it exists only in the deep Web. Future research is required to more accurately measure A-D.

Conclusion

Of the 3.3M unique web resources described in the Dublin Core metadata available through OAI-PMH repositories, approximately 700K (21%) were not indexed by any search engine. Yahoo indexed the most (65%), followed by Google (44%) and MSN (7%). Previous studies have estimated that Google has indexed more of the Web than other search engines, but we surmise that Yahoo scored the best in this study because of their agreement with OAIster. To date, not all OAI-PMH repositories register with one of the 4 popular registries, most likely because registration is optional and there is little perceived benefit of registering. However, if the popular search engines were to directly support OAI-PMH (and not indirectly through intermediaries), we believe the interest in registering and implementing OAI-PMH repositories would increase. Search engines would benefit by being able to index more content, and DLs would benefit by being able to share their contents with search engines without incurring web crawling overhead.

References

[1] M. K. Bergman, "The Deep Web: Surfacing Hidden Value," Journal of Electronic Publishing, August 2001, vol. 7, issue 1, http://www.press.umich.edu/jep/07-01/bergman.html.

[2] A. Gulli and A. Signorini, "The Indexable Web is More than 11.5 Billion Pages," Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 902-903.

[3] N. Paskin, "E-citations: Actionable Identifiers and Scholarly Referencing," Learned Publishing, vol. 13, no. 3, 2002, pp. 159-168.

[4] K. Shafer, S. Weibel, E. Jul, and J. Fausey, Persistent Uniform Resource Locators, http://www.purl.org

[5] S. Sun, L. Lannom, and B. Boesch, Handle System Overview. Internet Engineering Task Force (IETF), Request For Comments (RFC) 3650, November 2003.

[6] "U-M expands access to hidden electronic resources with OAIster," News Service, University of Michigan. http://www.umich.edu/news/index.html?Releases/2004/Mar04/r031004

URLs

1. http://www.openarchives.org/Register/BrowseSites
2. http://gita.grainger.uiuc.edu/registry/Info.asp
3. http://celestial.eprints.org/
4. http://archives.eprints.org/
5. http://www.nla.gov.au/digicoll/oai/

Frank McCown is a Ph.D. student in computer science at Old Dominion University. His research interests include digital preservation, web crawling, and web infrastructure. Contact him at fmccown@cs.odu.edu.

Xiaoming Liu is a technical staff member of the research library at Los Alamos National Laboratory. His research interests include repositories, digital preservation and web crawling. Contact him at liu_x@lanl.gov.

Michael L. Nelson is an assistant professor of computer science at Old Dominion University. His research interests include repositories, complex digital objects and digital preservation. He is a co-editor of the OAI-PMH. Contact him at mln@cs.odu.edu.

Mohammad Zubair is a professor of computer science at Old Dominion University. His research interests include digital libraries, information extraction and high performance computing. Contact him at zubair@cs.odu.edu.

OAI-PMH (side bar 1)

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) began as grass-roots interoperability effort for eprint archives [1]. It has since proven to be useful in a number of scenarios where loose synchronization of XML-encoded data is needed. OAI-PMH is not based on distributed searching. Instead, there are six "verbs" a *harvester* uses to request data from *repositories* [2]. Harvesters can then provide services (for either end-users or other harvesters) on the XML-encoded data they have collected.

The OAI-PMH has a simple, flexible data model. Owing to its eprints origins, the data model is generally interpreted in terms of bibliographic metadata describing a scholarly resource (Figure 1) although other interpretations are possible [3]. At the top of the model is the *resource* – the "thing" that is being described. This can be a traditional library object (e.g., book, report), or even non-digital entities (e.g., paintings, concepts). Next is the *item*, or the gateway to all metadata that describes the *resource*. The *item* provides a unique identifier with the metadata that describes the *resource*. Finally, at the bottom of the data model are the *records*. *Records* describe the *resource* in any metadata format that can be expressed as an XML Schema. Although OAI-PMH requires Dublin Core support as a lingua franca for cross-domain resource discovery, exporting richer metadata formats is encouraged (Figure 1). *Records* are uniquely identified by the *item* identifier, metadata format, and datestamp of creation or last modification (Figure 1 shows the scenario where datestamps are the same across all *records*). After an initial baseline harvest, a harvester can use the datestamp to request from a repository only those records that have changed since the previous harvest.

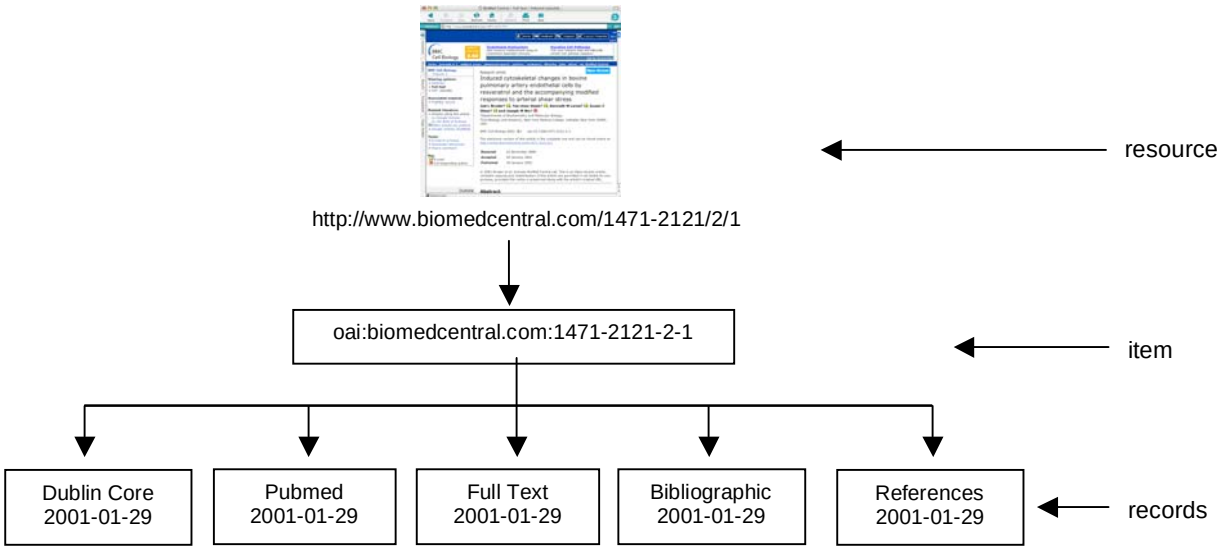

Figure 1. OAI-PMH Data Model

References

1. H. Van de Sompel and C. Lagoze, "The Santa Fe Convention of the Open Archives Initiative," D-Lib Magazine, vol. 6, num. 2, 2000.
2. C. Lagoze, H. Van de Sompel, M. L. Nelson, and S. Warner, The Open Archives Initiative Protocol for Metadata Harvesting, 2001. Available at http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
3. H. Van de Sompel, J. A. Young and T. B. Hickey, "Using the OAI-PMH… Differently," D-Lib Magazine, vol. 9, num. 7/8, 2003.

Harvesting the Deep Web (sidebar 2)

Harvesting the deep Web through issuing automating queries against search interfaces has been well documented [1 ,3, 5, 8, 9]. But recently interest has increased regarding having websites reveal their contents to web crawlers in a structured manner. Based on research first described in [2], Google's Sitemap Protocol allows a website to provide Google a set of URLs, their change rates, and their relative importance [4]. The mod_oai Apache module provides similar services but uses OAI-PMH to automatically generate and deliver website metadata and content [7].

Various methods have been implemented to allow search engines to collect data from OAI-PMH repositories. DP9 harvests records from OAI-compliant archives in batches and converts them into web pages [6]. The pages can then be crawled and indexed by search engines. The Extensible Repository Resource Locators (ERRoLs) for OAI Identifiers project allows the creation of URLs that dynamically perform OAI-PMH queries against registered OAI repositories and generate HTML pages suitable for web crawling [10].

OAI-PMH has been slowly making its way into the commercial search engines. Google supports OAI-PMH by allowing website operators to submit OAI-PMH baseURLs. Google is currently using OAI-PMH to index data from the National Library of Australia Digital Object Repository [URL 5]. Yahoo made an agreement with OAIster in March 2004 to acquire content harvested with OAI-PMH from 267 international research institutions. It is not known if Yahoo used OAI-PMH to obtain the content or if they used some other mechanism. It is also not known if Yahoo is continuing to obtain new OAIster content. OAIster has currently harvested more than 5.6 million records from 503 institutions. Many of the larger repositories used in our study were included in their list of contributing institutions.

References

1. A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In Proceedings from ACM SIGMOD, 2003, pp. 337-348.
2. O. Brandman, J. Cho, H. Garcia-Molina, and N. Shivakumar. Crawler-friendly web servers, SIGMETRICS Perform. Eval. Rev., vol. 28, num. 2, 2000, pp. 9-14.
3. V. Crescenzi, G. Mecca, and P. Merialdo. "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," VLDB Journal, 2001, pp. 109-118.
4. Google Sitemap Protocol, https://www.google.com/webmasters/sitemaps/docs/en/protocol.html.
5. P. G. Ipeirotis and L. Gravano, "Distributed search over the hidden-web: Hierarchical sampling and selection," In Proceedings of VLDB '02, 2002, pp. 394-405.
6. X. Liu, K. Maly, M. Zubair, and M.L. Nelson, "DP9: an OAI Gateway Service for Web Crawlers," In Proceedings of the Joint Conference on Digital Libraries (JCDL), June 2002, pp. 283-284.
7. M. L. Nelson, H. Van de Sompel, X. Liu, T. Harrison, N. McFarland, "mod_oai: An Apache Module for Metadata Harvesting," In Proceedings of ECDL 2005, Vienna, Austria, pp. 509-510.
8. A. Ntoulas, P. Zerfos, J. Cho, "Downloading Textual Hidden Web Content by Keyword Queries," In Proceedings of the Joint Conference on Digital Libraries (JCDL), June 2005, pp. 100-109.
9. S. Raghavan and H. Garcia-Molina, "Crawling the Hidden Web," In Proceedings of VLDB '01, 2001, pp. 129-138.
10. J. Young, Extensible Repository Resource Locators (ERRoLs) for OAI Identifiers, http://www.oclc.org/research/projects/oairesolver/.